Some Applications of Multiple Classes G-Networks with Restart

Jean Michel Fourneau[1(✉)] and Katinka Wolter[2]

[1] DAVID, UVSQ, Versailles, France
jean-michel.fourneau@uvsq.fr
[2] Frei Universitat, Berlin, Germany

Abstract. We show how to model system management tasks such as load-balancing and delayed download with backoff penalty using G-networks with restart. We use G-networks with a restart signal, multiple classes or positive customers, PS discipline and arbitrary PH service distribution. The restart signal models the possibility to abort a task and send it again after changing its class and its service distribution. These networks have been proved to have a product form steady-state distribution.

Keywords: Performance · G-Networks · Phase-type distributions · Product form steady-state distribution · Restart

1 Introduction

Since the seminal papers [2,5,6] published by Gelenbe more than 20 years ago, G-networks of queues have received considerable attention. G-networks have been previously presented to model Random Neural Networks [7,8]. They contain queues, customers (like ordinary networks of queues) and signals which interact with the queues and disappear instantaneously. Due to these signals G-networks exhibit much more complex synchronization and allow to model new classes of systems (artificial or biological). Despite this complexity, most of the G-networks studied so far have a closed form solution for their steady-state.

For most of the results already known, the effect of the signal is the cancelation of customer or potential (for an artificial random neuron) [1]. Recently, we have studied G-networks with multiple classes where the signal is used to change the class of a customer in the queue [4]. Such a signal is denoted as a restart because in some models it is used to represent that a task is aborted and submitted again (i.e. restarted) when it encounters some problems (see [9,10] for some systems with restart). These models still have a product form steady-state solution under some technical conditions on the queue loads.

Here we present some examples to illustrate how this new model and theoretical result can help to evaluate the performance of a complex system. We hope that this result and the examples presented here open new avenues for research and applications of G-networks. The technical part of the paper is organized as follows. The model and the results proved in [4] are introduced in Sect. 2 while the examples are presented in Sect. 3.

T. Czachórski et al. (Eds.): ISCIS 2016, CCIS 659, pp. 126–133, 2016.
DOI: 10.1007/978-3-319-47217-1_14

2 Model Assumptions and Closed Form Solutions

We have considered in [4] generalized networks with an arbitrary number N of queues. We consider K classes of positive customers and only one class of signals. The external arrivals to the queues follow independent Poisson processes. The external arrival rate to queue i is denoted by $\lambda_i^{(k)}$ for positive customers of class k and Λ_i^- for signals. The customers are served according to the processor sharing (PS) policy. The service times are assumed to be Phase-type distributed, with one input (say 1) and one output state (say 0). At phase p, the intensity of service for customers of class k in queue i is denoted as $\mu_i^{(k,p)}$. The transition probability matrix $H_i^{(k)}$ describes how, at queue i, the phase of a customer of class k evolves. Thus the service in queue i is an excursion from state 1 to state 0 following matrix $H_i^{(k)}$ for a customer of class k. We consider a limited version of G-networks where the customers do not change into signals at the completion of a service. Here, customers may change class while they move between queues but they do not become signals. More precisely, a customer of class k at the completion of its service in queue i may join queue j as a customer of class l with probability $P_{i,j}^{+(k,l)}$. It may also leave the network with probability $d_i^{(k)}$. We assume that a customer cannot return to the queue it has just left: $P_{i,i}^{+(k,l)} = 0$ for all i, k and l. As usual, we have for all i,k: $\sum_{j=1}^{N} \sum_{l=1}^{K} P_{i,j}^{+(k,l)} + d_i^{(k)} = 1$.

Signals arrive from the outside according to a Poisson process of rate Λ_i^- at queue i. Signals do not stay in the network. Upon its arrival into a queue, a signal first choses a customer, then it interacts with the selected customer, and it finally vanishes instantaneously. If, upon its arrival, the queue is already empty, the signal also disappears instantaneously without any effect on the queue. The selection of the customer is performed according to a random distribution which mimics the PS scheduling. At state $\boldsymbol{x}_i$, the probability for a customer to be selected is $\frac{x_i^{(k,p)}}{|\boldsymbol{x}_i|} \mathbb{1}_{\{|\boldsymbol{x}_i|>0\}}$ and the signal has an effect with probability $\alpha_i^{(k,p)}$. The effect is the restarting of the customer: this customer (remember it has class k and phase p) is routed as a customer of class l at phase 1 with probability $R_i^{(k,l)}$. We assume for all k, $R_i^{(k,k)} = 0$. Of course we have for all k, $\sum_{l=1}^{K} R_i^{(k,l)} = 1$ (Fig. 1).

The state of the queueing network is represented by the vector $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N)$, where the component $\boldsymbol{x}_i$ denotes the state of queue i. As usual with multiple class PS queues with Markovian distribution of service, the state of queue i is given by the vector $(x_i^{(k,p)})$, for all class indices k and phase indices p. Clearly $\boldsymbol{x}$ is a Markov chain. Let us denote by $|\boldsymbol{x}_i|$ the total number of customers in queue i. In [4] we have proved that the steady-state distribution, when it exists, has a product-form solution under some technical conditions on a fixed point system on the load.

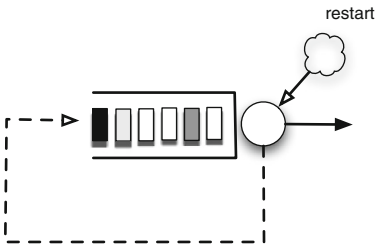

Fig. 1. Model of a queue with restart. The colors represent the classes

Theorem 1. *Consider an arbitrary open G-network with p classes of positive customers and a single class of negative customers the effect of which is to restart one customer in the queue. If the system of linear equations:*

$$\rho_i^{(k,1)} = \frac{\lambda_i^{(k)} + \sum_{o=1}^{P} \mu_i^{(k,o)} \rho_i^{(k,o)} H_i^{(k)}[o,1] + \nabla_i^{k,1} + \Delta_i^{k,1}}{\mu_i^{(k,1)} + \Lambda_i^- \alpha_i^{(k,1)}}, \tag{1}$$

where

$$\Delta_i^{k,1} = \sum_{p=1}^{P} \sum_{l=1}^{K} \Lambda_i^- \alpha_i^{(l,p)} \rho_i^{(l,p)} R_i^{(l,k)}, \tag{2}$$

$$\nabla_i^{k,1} = \sum_{j=1}^{N} \sum_{l=1}^{K} \sum_{q=1}^{P} \mu_j^{(l,q)} \rho_j^{(l,q)} H_j^{(l)}[q,0] P_{j,i}^{+(l,k)}, \tag{3}$$

and,

$$\forall p > 1, \quad \rho_i^{(k,p)} = \frac{\sum_{o=1}^{P} \mu_i^{(k,o)} \rho_i^{(k,o)} H_i^{(k)}[o,p]}{\mu_i^{(k,p)} + \Lambda_i^- \alpha_i^{(k,p)}} \tag{4}$$

has a positive solution such that for all stations i $\sum_{k=1}^{K} \sum_{p=1}^{P} \rho_i^{(k,p)} < 1$*, then the system stationary distribution exists and has product form:*

$$p(\boldsymbol{x}) = \prod_{i=1}^{N} (1 - \sum_{k=1}^{K} \sum_{p=1}^{P} \rho_i^{(k,p)}) |\boldsymbol{x}_i|! \prod_{k=1}^{K} \prod_{p=1}^{P} \frac{(\rho_i^{(k,p)})^{x_i^{(k,p)}}}{x_i^{(k,p)}!}. \tag{5}$$

Property 1. *This result is used to obtain closed form solutions for some performance measures: the probability to have exactly m customers in the queue and the expected number of customers in the queue.*

$$Pr(m\ customers) = (1 - \sum_{k=1}^{K}\sum_{p=1}^{P}\rho_i^{(k,p)}) \left[\sum_{k=1}^{K}\sum_{p=1}^{P}\rho_i^{(k,p)}\right]^m,$$

$$E[N] = \frac{\sum_{k=1}^{K}\sum_{p=1}^{P}\rho_i^{(k,p)}}{1 - \sum_{k=1}^{K}\sum_{p=1}^{P}\rho_i^{(k,p)}}. \tag{6}$$

3 Examples

We now present some examples to put more emphasis on the modeling capabilities of G-networks with restart signals. We model a load balancing system where the restarts are used to migrate the customers between queues and a back off mechanism for delayed downloading.

Example 1. Load Balancing: We consider two queues in parallel as depicted in Fig. 2. We want to represent a load balancing mechanism between them and we want to get the optimal rates to operate this mechanism and obtain the best performance.

The queues receive two types of customers: type 1 customers need to be served while type 2 customers represent the customers which must be moved to the other queue to balance the load. Customers of type 1 arrive from the outside according to two independent Poisson process with rate $\lambda_1^{(1)}$ for queue 1 and $\lambda_2^{(1)}$ for queue 2. There are no arrivals from the outside for type 2 customers. Type 2 customers are created by a restart. The service rates do not depend on the queue. They are equal to $\mu^{(1)}$ for type 1 and $\mu^{(2)}$ for type 2. For the sake of simplicity, we assume here that the service distributions are exponential. PH distributions will be added at the end of this example.

Restarting signals arrive to queues 1 and 2 according to two independent Poisson processes with rate Λ_1^- and Λ_2^-. When it arrives to a queue, a signal choses a customer at random as mentioned in the previous section and tries to change it to type 2. We assume the following probabilities of success: $\alpha_1^{(1)} = 1$ and $\alpha_1^{(2)} = 0$. Similarly, $\alpha_2^{(1)} = 1$ and $\alpha_2^{(2)} = 0$. Note that we have simplified the notation as we only have one phase of service (we consider exponential rather than PH distributions). This value of the acceptance probability means that the restarting signals is always accepted when the signal selects a type 1 customer and it fails when it tries to restart a type 2 customer (as by definition in this model, a type 2 customer is already restarted).

After its service, a type 1 customer leaves the system while a type 2 customer moves to the other queue and changes its type during the movement to become a type 1 customer. Thus the load balancing mechanism proceeds as follows: the signal is received by the queue and it selects a customer at random. If the customer has type 2, nothing happens. If the selected customer has type 1, it is restarted as a type 2 customer with another service time distribution and another routing matrix. The service time for a type 2 customer represents the

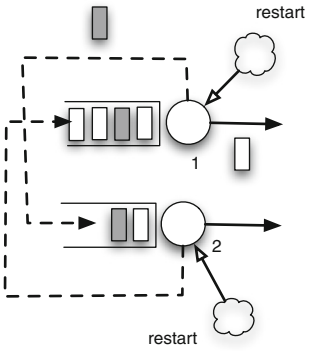

Fig. 2. Two queues in parallel with load balancing performed by restart signals

time needed to organize the job migration. It is assumed that it is much shorter than the the service type of a type 1 customer which represents the effective service. Let us now write the flow equations:

$$\rho_1^{(1)} = \frac{\lambda_1^{(1)} + \rho_2^{(2)}\mu^{(2)}}{\mu^{(1)} + \Lambda_1^-}, \quad \rho_1^{(2)} = \frac{\Lambda_1^- \rho_1^{(1)}}{\mu^{(2)}}, \quad \rho_2^{(1)} = \frac{\lambda_2^{(1)} + \rho_1^{(2)}\mu^{(2)}}{\mu^{(1)} + \Lambda_2^-}, \; \rho_2^{(2)} = \frac{\Lambda_2^- \rho_2^{(1)}}{\mu^{(2)}}. \tag{7}$$

Let us now consider the performance of such a system. We control the system with the rate of arrival of signals Λ_1^- and Λ_2^- and the objective is to balance the load with the smallest overhead. More formally, we say that the system is balanced if the loads for customers in service (i.e. not preparing their migration) are equal for both queues (i.e. $\rho_1^{(1)} = \rho_2^{(1)} = \rho$) and we assume that the overhead is the load of the queues due to the migration (i.e. $\rho_1^{(2)} + \rho_2^{(2)}$). Assuming that the system is balanced, we have:

$$\rho = \frac{\lambda_1^{(1)} + \rho_2^{(2)}\mu^{(2)}}{\mu^{(1)} + \Lambda_1^-} = \frac{\lambda_2^{(1)} + \rho_1^{(2)}\mu^{(2)}}{\mu^{(1)} + \Lambda_2^-}$$

After substitution, we get: $\rho = \frac{\lambda_1^{(1)} + \rho\Lambda_2^-}{\mu^{(1)} + \Lambda_1^-} = \frac{\lambda_2^{(1)} + \rho\Lambda_1^-}{\mu^{(1)} + \Lambda_2^-}$. Without loss of generality we assume that $\lambda_1^{(1)} > \lambda_2^{(1)}$. Taking into account the first part of the equation, we obtain: $\rho(\Lambda_1^- - \Lambda_2^-) = \lambda_1^{(1)} - \rho\mu^{(1)}$. Similarly using the second equation we get:

$$\rho(\Lambda_1^- - \Lambda_2^-) = \rho\mu^{(1)} + \lambda_2^{(1)}.$$

Thus, $\rho = \frac{\lambda_1^{(1)} - \lambda_2^{(1)}}{2\mu^{(1)}}$, and $\Lambda_1^- - \Lambda_2^- = \frac{\lambda_1^{(1)} + \lambda_2^{(1)}}{2}$. Taking now the other part of the objective into account we want to minimize the overhead of the load balancing mechanism. Remember that the global overhead is:

$$\rho_1^{(2)} + \rho_2^{(2)} = \rho\frac{(\Lambda_1^- + \Lambda_2^-)}{\mu^{(2)}}.$$

Thus the optimal solution is achieved for $\Lambda_2^- = 0$ and $\Lambda_1^- = \frac{\lambda_1^{(1)}+\lambda_2^{(1)}}{2}$. Let us now consider a more complex problem where the services for type 1 customer follow the same PH distribution. We still assume that type 2 customers receive services with an exponential distribution. Let us now write the flow equations:

$$\begin{aligned}
\rho_1^{(1,1)} &= \frac{\lambda_1^{(1)}+\sum_{p>0}\rho_2^{(2,p)}\mu^{(2,p)}}{\mu^{(1,1)}+\Lambda_1^-}, \quad \rho_1^{(1,p)} = \frac{H(1,p)\rho_1^{(1,1)}\mu^{(1,1)}}{\mu^{(1,p)}+\Lambda_1^-}, \forall p > 1, \\
\rho_2^{(1,1)} &= \frac{\lambda_2^{(1)}+\sum_{p>0}\rho_1^{(2,p)}\mu^{(2,p)}}{\mu^{(2,1)}+\Lambda_2^-}, \quad \rho_2^{(1,p)} = \frac{H(1,p)\rho_2^{(2,1)}\mu^{(2,1)}}{\mu^{(2,p)}+\Lambda_2^-}, \forall p > 1, \\
\rho_1^{(2)} &= \frac{\Lambda_1^- \sum_{p>0}\rho_1^{(1,p)}}{\mu^{(2)}}, \qquad \rho_2^{(2)} = \frac{\Lambda_2^- \sum_{p>0}\rho_2^{(1,p)}}{\mu^{(2)}}.
\end{aligned} \tag{8}$$

These equations can be used to optimize the system as we have done previously for exponential service distributions.

Example 2. Delayed Downloading: We now study a small wifi network with a delayed downloading mechanism (see for instance [11]). Queue A is the downloading queue (see Fig. 3). Customers and signals arrive from the outside to queue A. The class of customers represents the delays that requests will experience. Type 1 requests (in white) are not delayed while delayed requests are depicted in grey. The restart signals change the state of a request to "delayed" according to the selection mechanism described in Sect. 2. The probability of acceptance for the selection depends on the class of the customer and the phase of service. Thus, we can model delay based on the steps of the downloading protocol, for instance. Once a request class has been changed due to selection by the signal, it is routed after its service to queue B or C where it is changed again to a class 1 request and experiences a random delay depending on the queue. The flow equations are:

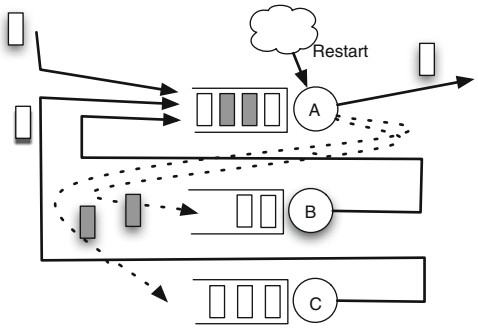

Fig. 3. The queuing network associated to the delayed downloading with back-off penalties

$$\rho_A^{1,1} = \frac{\sum_{o=1}^{P} \mu_A^{1,o}\rho_A^{1,o}H_A^{(k)}[o,1] \; + \; \sum_{p=1}^{P} \mu_B^{1,p}\rho_B^{1,p}H_B^{(1)}[p,0] \; + \; \sum_{p=1}^{P} \mu_C^{1,p}\rho_C^{1,p}H_C^{(1)}[p,0]}{\mu_A^{1,1} + \Lambda_A^- \alpha_A^{1,1}}, \tag{9}$$

$$\forall p > 1, \quad \rho_A^{1,p} = \frac{\sum_{o=1}^{P} \mu_A^{1,o}\rho_A^{1,o}H_A^{(1)}[o,p]}{\mu_A^{1,p} + \Lambda_A^- \alpha_A^{1,p}}, \quad and \;\; \forall k > 1, \rho_A^{k,p} = \frac{\sum_{o=1}^{P} \mu_A^{k,o}\rho_A^{k,o}H_A^{(k)}[o,p]}{\mu_A^{k,p}}, \tag{10}$$

$$\rho_A^{k,1} = \frac{\sum_{o=1}^{P} \mu_A^{k,o}\rho_A^{k,o}H_A^{(k)}[o,1] \; + \; \sum_{p=1}^{P} \Lambda_A^- \alpha_A^{(1,p)}\rho_A^{(1,p)}R_A^{(1,k)}}{\mu_A^{1,1} + \Lambda_A^- \alpha_A^{1,1}}, \tag{11}$$

$$\rho_B^{1,1} = \frac{\sum_{o=1}^{P} \mu_B^{1,o}\rho_B^{1,o}H_B^{(k)}[o,p] \; + \; \sum_{p=1}^{P} \mu_A^{2,p}\rho_A^{2,p}H_A^{(2)}[p,0]}{\mu_B^{1,1}}, \tag{12}$$

$$\forall p > 1, \quad \rho_B^{1,p} = \frac{\sum_{o=1}^{P} \mu_B^{1,o}\rho_B^{1,o}H_B^{(k)}[o,1]}{\mu_B^{1,p}}, \quad and \;\; \rho_C^{1,p} = \frac{\sum_{o=1}^{P} \mu_C^{1,o}\rho_C^{1,o}H_C^{(k)}[o,1]}{\mu_C^{1,p}}, \tag{13}$$

$$\rho_C^{1,1} = \frac{\sum_{o=1}^{P} \mu_C^{1,o}\rho_C^{1,o}H_C^{(k)}[o,p] \; + \; \sum_{p=1}^{P} \mu_A^{3,p}\rho_A^{3,p}H_A^{(3)}[p,0]}{\mu_C^{1,1}}. \tag{14}$$

Assuming that these equations have a fixed point solution such that the queues are stable, Theorem 1 proves that the steady-state distribution has product form. This closed form solution allows us to study the performance of the downloading mechanism and to optimize the throughput when one changes the delay distributions.

4 Concluding Remarks

Note that it is possible to add triggers in the model to increase the flexibility while conserving the closed form solution [3]. We advocate that G-networks with restart signals are a promising and flexible modeling technique.

Acknowledgments. This work was partially supported by project MARMOTE (ANR-12-MONU-00019) and by a PROCOPE PHC grant between Université de Versailles and Frei Universitat, Berlin.

References

1. Artalejo, J.R.: G-networks: a versatile approach for work removal in queuing networks. European J. Op. Res. **126**, 233–249 (2000)
2. Fourneau, J.M., Gelenbe, E., Suros, R.: G-networks with multiple classes of positive and negative customers. Theor. Comput. Sci. **155**, 141–156 (1996)
3. Fourneau, J.-M., Wolter, K.: Mixed networks with multiple classes of customers and restart. In: Remke, A., Manini, D., Gribaudo, M. (eds.) ASMTA 2015. LNCS, vol. 9081, pp. 73–86. Springer, Heidelberg (2015)
4. Fourneau, J.M., Wolter, K., Reinecke, P., Krauß, T., Danilkina, A.: Multiple class G-networks with restart. In: ACM/SPEC International Conference on Performance Engineering, ICPE 2013, pp. 39–50. ACM (2013)
5. Gelenbe, E.: Product-form queuing networks with negative and positive customers. J. Appl. Probab. **28**, 656–663 (1991)
6. Gelenbe, E.: G-networks with instantaneous customer movement. J. Appl. Probab. **30**(3), 742–748 (1993)
7. Gelenbe, E.: G-networks: an unifying model for queuing networks and neural networks. Ann. Oper. Res. **48**(1–4), 433–461 (1994)
8. Gelenbe, E., Fourneau, J.M.: Random neural networks with multiple classes of signals. Neural Comput. **11**(4), 953–963 (1999)
9. van Moorsel, A.P.A., Wolter, K.: Analysis and algorithms for restart. In: 1st International Conference on Quantitative Evaluation of Systems (QEST 2004), The Netherlands, pp. 195–204. IEEE Computer Society (2004)
10. van Moorsel, A.P.A., Wolter, K.: Analysis of restart mechanisms in software systems. IEEE Trans. Softw. Eng. **32**(8), 547–558 (2006)
11. Wu, H., Wolter, K.: Analysis of the energy-performance tradeoff for delayed mobile offloading. In: Proceedings of the 9th EAI International Conference on Performance Evaluation Methodologies and Tools, VALUETOOLS 2015, pp. 250–258 (2015)